NOTES AND LITERATURE

BIOMETRICS

An Important Contribution to Statistical Theory

One of Pearson's most valuable contributions to statistical theory is his test for goodness of fit.[1] It enables one, with the aid of Elderton's[2] tables, easily to determine the probability that a given system of observed frequencies does or does not differ significantly from a series of theoretical frequencies supposed to graduate the observations. The significance of this criterion in Mendelian work has recently been pointed out by Harris.[3]

Hitherto this criterion has found an important limitation in the fact that, as originally developed by Pearson, it was applicable only to frequency systems. It could be used to test goodness of fit only where the observations were counts of the number of times particular classes of events occurred. But, of course, frequency systems comprise only one kind of observational data to which one has occasion to fit curves. Much more often there is need for a criterion of goodness of fit where the observations are of the nature of true ordinates, rather than frequencies. Such cases include all data of the sort where a mean y is determined for each x, as in a growth curve; or in the regression observed in a correlation table, where for each successive value of one of the variables the mean value of the correlated variable is calculated. There has been no method of testing the goodness of fit for such curves. From a visual inspection of the plotted regression line one has been compelled to form his judgment as to whether it was or was not a good fit.

Recently a Russian statistician, E. Slutsky,[4] has extended

1 Pearson, K., "On the Criterion that a Given System of Deviation from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to Have Arisen from Random Sampling," *Phil. Mag.*, 5th Series, Vol. L, pp. 157–175, 1900.

2 *Biometrika,* Vol. I, pp. 155–163.

3 Harris, J. A., "A Simple Test of the Goodness of Fit of Mendelian Ratios," Amer. Nat., Vol. 46, 1912, pp. 741–745, 1912.

4 Slutsky, E., "On the Criterion of Goodness of Fit of the Regression Lines and on the Best Method of Fitting Them to the Data," *Jour. Roy. Stat. Soc.,* Vol. LXXVII, Part I (December, 1913), issued 1914, pp. 78–84.

Pearson's theory to cover the class of curves, formerly not amenable to such test. The result forms an extremely valuable extension of biometric theory.

Briefly Slutsky's essential result may be put as follows. He finds (the complete proof is not given in this paper) that

$$\chi^2 = S\left(\frac{n_{x_p} e_p^2}{\sigma_{n_{x_p}}^2}\right),$$

where χ^2 is the quantity denoted by the same letter in Pearson's original work, and is the argument in Elderton's table; n_{x_p} is the frequency in the x_p array, *i. e.*, the number of observations on which each observed ordinate is based; e_p is the difference between the observed and the calculated mean y for each x_p array; and $\sigma_{n_{x_p}}$ is the standard deviation of each x_p array; *i. e.*, the standard deviation of the group of observations from which each particular y was calculated. S, as usual, denotes summation. Knowing χ^2, P is read directly from Elderton's tables.

Slutsky gives a couple of examples of the application of the method in his paper. For illustration here I have preferred to take an example from my own unpublished data. The observations (y_{x_p}) in this case are the mean butter productions of American Jersey cattle, based on seven-day tests.[5]

The theoretical points Y_{x_p} are calculated from the equation,

$$y = 14.21098 + .0250x - .0038x^2 + 3.0104 \log x,$$

the constants of which were determined from the observations by the method of least squares.

The test for goodness of fit is carried out in Table I. It should be said that, following the suggestion given by Slutsky in his paper, I have used in the $\sigma_{n_{x_p}}$ column the graduated rather than the observed values. In the present case the scedastic curve is hopelessly far from a straight line. It is, in point of fact, logarithmic.

From this table we have $\chi^2 = 32.115$. This is beyond the range of Elderton's table. By a rough, but sufficiently accurate, graphical extrapolation, I find for present values of n' and χ^2,

$$P = .417 \text{ about.}$$

In other words, if the butter production of Jersey cows changes with age according to the curve given, we should expect to

[5] For data see "Jersey Sires and Their Tested Daughters," published by American Jersey Cattle Club, New York, 1909.

get a worse agreement between observation and theory in 42 out of every 100 random samples on which the point was tested. In other words, the fit may be considered sufficiently good. As a *matter of fact, the fit is extraordinarily close over most of the* curve. Four (only) out of the 32 ordinates contribute more than 50 per cent. of the value of χ^2.

TABLE I

Age in Years	Observed Butter Production in Lbs.	Calc. Butter Production in Lbs.	Errors	Frequency	Standard Dev. of Arrays	
x_p	y_{x_p}	Y_{x_p}	$e_p = (Y_{x_p} - y_{x_p})$	n_{x_p}	$\sigma_{n_{x_p}}$	$\frac{n_{x_p} e^2_p}{\sigma^2_{n_{x_p}}}$
1.25	14.25	14.23	.02	2	.04	.500
1.75	15.15	15.15	.00	46	.97	.000
2.25	15.57	15.69	.12	273	1.49	1.771
2.75	15.96	16.06	.10	312	1.83	.932
3.25	16.38	16.35	.03	545	2.07	.114
3.75	16.72	16.57	.15	511	2.25	2.271
4.25	16.92	16.74	.18	704	2.38	4.027
4.75	17.09	16.89	.20	532	2.49	3.432
5.25	17.01	17.00	.01	556	2.56	.008
5.75	17.07	17.09	.02	382	2.62	.022
6.25	16.98	17.16	.18	419	2.65	1.933
6.75	17.04	17.21	.17	277	2.68	1.114
7.25	17.09	17.25	.16	285	2.68	1.016
7.75	17.48	17.27	.21	190	2.68	1.167
8.25	17.30	17.28	.02	166	2.67	.009
8.75	17.17	17.27	.10	121	2.64	.174
9.25	17.56	17.25	.31	109	2.61	1.515
9.75	16.67	17.21	.54	95	2.57	4.194
10.25	17.05	17.17	.12	63	2.52	.143
10.75	17.42	17.11	.31	39	2.46	.619
11.25	16.95	17.05	.10	54	2.40	.094
11.75	17.00	16.97	.03	28	2.33	.005
12.25	17.05	16.88	.17	20	2.26	.113
12.75	16.54	16.79	.25	7	2.18	.092
13.25	16.34	16.68	.34	11	2.09	.291
13.75	18.14	16.56	1.58	9	1.99	5.673
14.25	15.89	16.44	.55	7	1.88	.599
14.75	16.15	16.30	.15	5	1.77	.036
15.25	16.37	16.16	.21	4	1.65	.065
15.75	15.75	16.00	.25	2	1.53	.053
16.25	15.42	15.84	.42	3	1.40	.117
16.75	15.75	15.67	.08	4	1.27	.016
Totals . . .				5,781		32.115

It may be said, in conclusion, that Slutsky's contribution is one which will be highly valued by all investigators who have a critical interest in the graduation of observational data, whatever the field in which they may be working.

RAYMOND PEARL